Cultural influences on word meanings revealed through large-scale semantic alignment

Bill Thompson,[1,*] Seán G. Roberts,[2,3] Gary Lupyan[4]

[1]Department of Computer Science, Princeton University, Princeton, NJ, USA
[2]School of English, Communication and Philosophy, Cardiff University, Cardiff, UK
[3]Department of Anthropology and Archaeology, University of Bristol, Bristol, UK
[4]Department of Psychology, University of Wisconsin, Madison, WI, USA

[*]To whom correspondence should be addressed; E-mail: wdt@princeton.edu

If the structure of language vocabularies mirrors the structure of natural divisions that are universally perceived, then the meanings of words in different languages should closely align. In contrast, if shared word meanings are a product of shared culture, history, and geography, they may differ between languages in substantial but predictable ways. We analyzed the semantic neighbourhoods of 1,010 meanings in 41 languages. The most aligned words were from semantic domains with high internal structure (number, quantity, kinship). Words denoting natural kinds, common actions, and artifacts aligned much less well. Languages that are more geographically proximate, more historically related, and languages spoken by more similar cultures, had more aligned word meanings. These results provide evidence that the meanings of common words vary in ways that reflect the culture, history and geography of their users.

Everyday words like "red", "sad", "house", "run", and "sister" may strike us as denoting concepts that exist independently of any language. On a traditional view, words like these map onto conceptual categories that we acquire independently of experience with any language.[1–4] On a strong version of this universalist view, word meanings are “more-or-less straightforward mappings from a preexisting conceptual space, programmed into our biological nature: humans invent words that label their concepts.”[5] Alternatively, vocabularies of different languages may reflect different solutions to categorizing objects, relations, actions, and abstract ideas.[6–9] On this alternative, relative, perspective, language vocabularies are culturally evolved sets of categories that we learn in the course of learning a language.[10] Rather than reflecting an innate store of concepts, or simply mapping onto categories extracted by a common perceptual system, “the categories and types that we isolate from the world ... we do not find there because they stare every observer in the face, [but because they are organized by] linguistic systems...”[11]

The universalist and relative perspectives make some of the same predictions: Both predict that languages may have many “culture-bound”[12] words that have no translation-equivalents in another language. It would surprise no one to discover that regional animals and natural features (“kangaroo”, “fjord”), specialized artifacts (‘carburetor”), technical terms (“methylation”), and complex social constructs (“sabbath”) may be absent from certain languages. It is precisely because of the non-universality of such meanings that languages tend to borrow words that denote them wholesale from other languages.[13,14] Both perspectives likewise allow vocabularies to adapt to differences in communicative need.[15] To the extent that people in colder climates are more likely to need to distinguish between “ice” and “snow”, we ought to find that languages spoken in colder climates are more likely to lexicalize this difference, and indeed we do,[16] see also.[4,17–19]

However, when it comes to common “everyday” meanings, the predictions of the universalist and relative perspectives diverge. The universalist perspective predicts that words denoting common animals and artifacts, common natural features (e.g., "river", "sand"), basic emotions, body parts, and common actions — meanings that ought to be similarly available to everyone regardless of the language they speak — should closely align across languages. On the whole, concrete terms may be expected to vary less (i.e., align better) than abstract terms. Differences, where they are found, should be random and unpredictable. In contrast, the relative view (in agreement with the intuitions of many lexicographers[12,20]) predicts that words denoting even highly concrete and seemingly self-evident meanings may fail to align across languages. Importantly, the degree of alignment should be predictable from cultural, historical, and geographic factors. Languages that are geographically closer, have more recent common ancestors, and are spoken by more culturally similar groups, should have words that are more alignable in meaning.

In the analyses that follow, we examine which semantic domains (e.g., animals, emotions, body parts, numbers) show most and least alignment between different languages, and whether alignment is greater for more concrete terms, as predicted by the universalist view. We then examine how alignment varies for different parts of speech and how alignment relates to lexical factors such as frequency and neighborhood density. Lastly, we examine whether the alignment between one language and another is related to cultural and historical relatedness of the two languages.

What does it mean for two words to mean “the same thing”? Semantic equivalence can be defined in functional terms: the meaning of a word w_1 in one language (L_1) is aligned with a word w_2 in another language (L_2) if the two words are used in the same contexts by L_1 and L_2 speakers. One reason why describing the semantic structure of natural languages is difficult is that word meanings, like other psychological constructs, are not directly observable.[21] The most direct way to assess semantic equivalence would be to query L_1-L_2 speakers of multiple languages about the meanings of different words.[22–26] For example, the English word “impressed” has unambiguously positive valence[27] while the valence of its Italian translation equivalent, "impressionato", is relatively more negative.[28] This difference in valence suggests that "impressed" and "impressionato" do not quite mean the same thing. This approach, however, is difficult to implement at scale. For this reason, existing attempts to quantify semantic structure have focused on either comprehensive analyses of a specific language (often English;[29,30]), cross-linguistic comparisons of a small set of meanings[31–35] or focus on a single domain such as emotion words.[36]

Here, we present a large-scale analysis of vocabularies spanning 1,010 distinct "concepts" in 41 languages (the concepts are listed in Supplementary Table 1.4.2.3). Our analysis specifically focuses on words that are, in principle, highly translatable and takes advantage of recent advances in distributional semantics. Distributional semantics is premised on the idea that it is possible to understand the meaning of words by observing the contexts in which they are used: "you shall know a word by the company it keeps".[37] The meanings of w_1 and w_2 are similar to the extent that the contexts in which they are used are similar. Early attempts to quantify semantic similarity using only the contextual information were surprisingly successful at learning reasonable semantic embeddings,[38–40] but were hampered by computational intractability. Advances in machine-learning,[41] combined with the availability of large corpora of digitized text, have now made it possible to estimate representations of word meanings – word embeddings - in a way that correlates with human semantic judgments with a surprising degree of subtlety.[42–51]

Semantic representations derived from word embeddings capture both the range of contexts in which a word is used and the relative frequencies of those contexts. Comparing contexts of use across languages allows us to quantify in a data-driven way what is sometimes called "partial equivalence"[52] – similarities and differences in the semantic profiles of translation-equivalent pairs. If word meanings reflect self-evident natural partitions or universal constraints on how people form concepts, we should find substantial regularities in these semantic profiles across languages. For example, if the English meanings of “in” and “out” depend on categories that are embedded in the physical world, e.g., inward motion/outward motion as perceived by all humans, then translation equivalents of the English “in” and “out” are likely to be used in all (or most) of the same contexts, yielding high alignment between these terms and their translation equivalents in other languages.

We obtained word-forms for 1,010 concepts in 41 languages using the NorthEuraLex (NEL) dataset.[53] NEL is compiled from dictionaries and other linguistic resources available for individual languages in Northern Eurasia. Translation pairs can be derived from NEL because it provides word forms for the same set of concepts in multiple languages. For example, NEL provides word forms for the concept “DOG” in 107 languages (e.g. English: “dog”, French: “chien”, Finish: “koira”). Each of the NEL concepts can be assigned to a semantic domain (for example, the concept “DOG” is assigned to the semantic domain “Animals”, while the concept “NOSE” is assigned to “The Body”) using the Concepticon organizing scheme (see Methods).

In our main analyses, we analyzed word embeddings derived from applying the fastText skipgram algorithm to language-specific versions of Wikipedia.[54] We also replicated these analyses using embeddings derived from the OpenSubtitles2018 database[55] and from a combination of Wikipedia and the Common Crawl dataset.[56] For details of these replications (SI 1.2.2) and others, including an analysis of alignment using a much larger set of translation equivalents (SI 1.2.1), and an analysis of how our alignment measure relates to alignment derived from patterns of colexification (SI 5.3.3); see Methods and Supplementary Information.

Figure 1 shows a schematic of our alignment algorithm (see also figure 2). For a given language pair (L_i and L_j) and concept (c), we first identified the closest k semantic neighbours of the word for c in the vector embeddings of L_i (restricted to words that can be translated into L_j; in our primary analyses, this means semantic neighbours are limited to the NEL vocabulary

– see SI 1.2.1 for analyses of larger translation vocabularies, and SI 1.1 for details of how our method deals with NEL concepts that are associated with multiple words). We then determined whether the translations of these neighbours are also close semantic associates of the word for c in L_j. The directional semantic alignment $L_i \rightarrow L_j$ is the Pearson correlation between the words for c's similarity to these neighbours in both languages. For example, in figure 2 the closest neighbours to the English word "beautiful" are "colourful" (.55), "love" (.53) and "delicate" (.51). French translations of these neighbours are more distant from "beau", ("multicolore"=.22, "aimer"=.32 and "fin'‘=.2). This reduces the correlation, so alignment is low in this direction (alignment is lowest when neighbor similarities are uncorrelated). The procedure was then repeated in the opposite direction: the k closest semantic neighbours to the word for c in L_j were identified and matched to their translations into L_i; the same Pearson correlation statistic was calculated for $L_j \rightarrow L_i$. The semantic alignment of c is the average of the two correlations. We will refer to this quantity as a. Alternative measures of alignment are discussed in the Supplementary Information (SI 1.4).

We used this algorithm to analyse semantic alignment in a dataset (see Methods) that includes 1,010 concepts across 21 semantic domains (e.g. Kinship, Animals, Body Parts) with an average of 48 concepts per domain (median: 40; min: 12; max: 136), in 781 language pairings from 41 languages, spanning 10 language families, with an average of 797 concepts per language pair (median: 837; min: 67; max: 991).

Results

Validating computed semantic alignment

Does lower semantic alignment correspond to words that mean different things in different languages? We validated that our alignment measure tracks differences in translatability in several ways. First, we obtained human-rated translation similarity for 201 Dutch-English translation pairs in our dataset.[24] Computed alignment was significantly correlated with Dutch-English translation similarity judgments (r=.33, p <.001). This moderate correlation increases to r=.60 (p=.02) when aggregated by the 15 semantic domains that contained ratings for at least 5 words, and remained a significant predictor when controlling for semantic domain (b = .14, 95% CI = [.078, .203], t = 4.37, p < .001) and differences in (log-transformed) word frequency (b = .13, 95% CI = [.068, .194], t = 4.08, p < .001). We further confirmed the positive relationship between computed semantic alignment and human ratings using a set of Japanese-English translatability ratings[26] for 192 word pairs. These too were significantly correlated with alignment (r=.29, p <.001); using an independent set of machine-generated translations to compute alignment achieved a nearly identical result (r=.30, p <.001).

As additional validation, we used our semantic alignment measure to predict differences in name agreement for 750 images each named by speakers of six languages (Spanish, English, French, Italian, German, and Netherlands Dutch).[57] Not surprisingly, some images are named more consistently (e.g., cat, gloves) than others (e.g., megaphone, clothes drying rack). We expected that meanings with lower semantic alignment will correspond to less consistent patterns of name agreement across the six languages. Overall, images with lower name-agreement (e.g., "[clothes] hanger", "gym") corresponded to words with lower overall alignment between these six languages (though the correlation is relatively small, r = .17, p <.001). More interestingly, while some images had high name agreement in all six languages, other images had high agreement in some languages, but not others. For example an image of a clothes hanger has high agreement in Spanish (100% produce "percha"), less in English (77% produce "hanger") and less still in Italian (only 33% produce the modal response, "appendino"). We predicted that such differences in agreement would be associated with lower alignment. Confirming this prediction, larger differences in name agreement were associated with lower alignment (b=-.20, 95% CI = [-.256, -.146], t = -7.21, p <.001). This relationship continued to remain reliable when adjusting for cross-linguistic differences in (log-transformed) word frequencies, and when taking into account the geographic and historical relationships between languages (b = -.13, 95% CI = [-.190, -.071], t = -4.28, $p < .001$).

Comparing alignment in 21 semantic domains

As shown in Fig. 3, alignment varied by semantic domain. On the universal perspective, alignment is predicted to be greatest for words denoting natural kinds and highly concrete meanings, such as common artifacts. Our analysis did not reveal support for this prediction. There was no statistically significant relationship between concreteness (derived from English-based norms[58]) and alignment ($t = -0.980, p = 0.33$). Some natural kind terms were relatively well aligned, e.g., "dog" (a = .37), "wind" (a = .38) "water" (a = .28). As a benchmark, we calculated the alignment of NEL concepts in English from two different corpora and found average alignment to be a = 0.53 (max = 0.98 for "thirty"; min = 0 for "rustle", see SI 1.3.2; see also SI 1.3.1 for further baseline analyses of cross-linguistic alignment). In light of this within-language expectation, terms like "dog" and "food" were well aligned across languages. However, other natural kind terms had among the lowest alignments: e.g. "feather" (a = .12), "branch" (a = .12). Similarly, words pertaining to universal aspects of human existence showed variability in alignment, e.g., "move" (a = .14), "sad (a = .32), and "food" (a = .42). The most aligned words were instead number words, temporal terms ("day", "week", "Spring"), and common kinship terms ("daughter", "son", "aunt") with alignments ranging between a = .49 and a = .84. Alignments for all 1,010 meanings are reported in Supplementary Table 1.4.2.3.

The meanings that are most alignable (e.g., numbers and kinship terms) stand out not as being especially concrete or reflecting "natural" joints, but as domains which have high internal coherence. Although kinship systems vary, terms denoting close kin relations are organized along a few dimensions such as gender (son/daughter, mother/father) and generation (grandmother/mother/daughter).[59,60] This low dimensionality seems to allow for high alignment. Similarly, although a base-10 counting system is a cultural invention, once adopted, it imposes strong constraints such that the (semantic) difference between the English words "five" and "ten" is nearly identical to the difference between the Spanish equivalents "cinco" and "diez". However, there is also systematic variation in number terms (see Fig. 4 and SI 4.2). Words for '1' and '2' have lower alignment than other numerals, likely because of being grammaticalized as indefinite and dual markers, respectively.[61] In general, alignment is lower for words with more polysemous meanings (e.g. Hungarian "hét" means '7' and 'week') – see SI 5.3.2 for an analysis of how semantic alignment relates to polysemy, as quantified using colexification networks.[36,62] Alignment is also lower for larger numbers (p <.001), possibly due to their lower absolute frequency in language.[63] For numbers of 50 or more, alignment is lower if the numbers are constructed with different numeral typologies (e.g. '80' in English is constructed as 8×10 but in French "quatre-vingts" is 4×20,,[32] interaction effect p <.001). These results are robust to controls for historical relatedness. Although our alignment measure is sensitive only to word co-occurrences, we can detect in the alignment patterns certain historical vestiges. For example, modern Danish uses a standard base-10 system but some number terms still reflect their historical roots in a base-20 system, (e.g., 60 "tres" = 3×20), and an archaic form of "half" (e.g., 70 "halvfjerds" = 3.5×20).[32] These irregular Danish number terms align significantly less well with the corresponding numeric terms in other languages compared to other Danish number terms (b = -.02, 95% CI = [-.025, -.018], t = -11.25, p < .001), see Fig. 4.

Predicting alignment from syntactic and lexical factors

We next examined how alignment related to several other lexical properties. Part-of-speech was a highly significant predictor of alignment, accounting for 16% of the variance ($p < .001$).

Verbs, Conjunctions, and Prepositions, were the least aligned; Wh-words and numerals were the most aligned [Fig. 5]. There was no statistically significant interaction between part of speech and concreteness ($\chi^2(8) = 1.58, p = .127$). Other significant predictors of alignment were absolute differences in word frequency and semantic neighborhood density[64] (a simple measure of the extent to which words are embedded in a system of semantically related terms). Larger differences in (log-transformed) word-frequencies[65] were correlated with lower alignment (r = -.20; b = -.04, 95% CI = [-.04, -.037], t = -77.25, $p < .001$). Likewise, greater differences in the (log-transformed) semantic neighborhood density (computed for all languages, see SI 1.1.5), were negatively correlated with alignment (r = -.13; b = -.01, 95% CI = [-.0104, -.0099], t = -67.72, $p < .001$). Semantic domain, part of speech, frequency and neighborhood density differences accounted for 30% of the variance in alignment in a mixed effects model with language-pair and concept as random effects.

Of further interest, our semantic alignment measure was strongly related to the rate at which word-forms change over time. How quickly a word-form changes is not only related to its frequency,[66] but also to its alignment. More aligned meanings tended to have word-forms that show slower rates of change (see SI 1.3.4).

Predicting semantic alignment from culture and history

The relative perspective predicts that languages spoken by people with more similar cultures should align to a greater extent. Confirming this prediction, we found that cultural similarity (the proportion of cultural traits in common based on 92 non-linguistic cultural traits for 39 societies representing 39 languages in our sample[67]) predicted semantic alignment between languages (b=.20, t=6.01, $p < .001$). Word meanings of more similar cultures aligned better. The same pattern was found for geographic distance (b= -.20, t=-6.42, $p <.001$), and for a patristic-distance-based measure (see SI 4.1) of language history (available only for Indo-

European languages, $b = -.178, t = -3.03, p = .002$). Cultural similarity continued to correlate with semantic alignment when controlling for language history and geographic proximity ($b = .25, t = 3.16, p = .002$). In these tests we used language families and geographic area as random effects to control for non-independence of languages. SI 4.1 presents additional tests that further assess the robustness of these relationships to non-independence.

Our finding that semantic alignment is predictable to a certain extent from culture, language history, and geography (R^2 = .363) contrasts with previous research based on patterns of polysemy, which failed to find these relationships and has been interpreted as support for the universalist perspective.[34] We calculated a polysemy-based alignment measure (see SI 5.2) using a recent, large-scale database[62] of common colexifications (an approach that has been successfully used to quantify semantic alignment in the specific semantic domain of emotion vocabulary, for example[36]). First, we established that the relationship between semantic alignment and cultural similarity is robust to controls for polysemy (see SI 5.3.3). Second, we examined whether colexification is related to cultural similarity (and geographic proximity) in the same way that our distributional measure of alignment is. This polysemy-based measure of semantic alignment was not statistically significantly related to cultural similarity, and was a much weaker predictor of geographic distance compared to the distributional measure of semantic alignment (see SI 4.1 and Discussion).

Using our distributional approach to alignment, we also investigated the relationship between overall cultural similarity and semantic alignment within each semantic domain (see SI 4.1). The strongest correlations were for words related to 'food and drink' (r = .29), 'time' (r = .27), 'animals' (r = .26), and 'the body' (r = .23; adjusted p-values $< .001$). The weakest correlations were for words related to 'motion', 'basic actions', 'emotion', and 'cognition' (adjusted p-values $> .1$). We can also compute cultural similarity for specific cultural domains (e.g., 'subsistence type', 'rituals', 'marriage and kinship'[67] rather than semantic domains for words. Cultural similarity related to 'subsistence type' was correlated with semantic alignment in domains including 'food and drink' (r = .30), 'animals' (r = .29), 'agriculture and vegetation' (r = .25), 'clothing and grooming' (r = .25), 'social and political relations' (r = .15), and 'spatial relations' (r = .10, all adjusted p-values < .05). These reflect well-known relations between subsistence types and culture.[18,68–75] Cultural similarity related to settlement (group size, community organisation etc.) was correlated with semantic alignment in domains including 'kinship' (r = .28), 'the physical world' (r = .13) and 'spatial relations' (r = .11, adjusted p-values < 0.05), also reflecting previous findings.[76,77] Finally, cultural similarity related to political organisation was correlated with the semantic alignment of words related to the body (r = .21, adjusted p = .001), perhaps reflecting the use of metaphors of society as a body.[78,79]

For 19 Indo-European languages where fine-grained historical and geographic proximity were available,[80,81] we found that semantic alignment was significantly correlated with historical proximity (r = .34, 95% CI = [.16,.5], one-tailed p = .01), but not geographic proximity (r = .26, 95% CI = [.18,.37], one-tailed p = .08). Figure 6 illustrates the semantic relationships between Indo-European languages inferred from their semantic alignments. For these languages, we estimated the relative contribution of geography, history and culture to alignment in each semantic domain. The relative effect of historical proximity did not differ much between domains. The relationship with cultural similarity was stronger than geographic proximity for most domains (3 largest differences were for kinship, animals and the body, 3 smallest differences were for motion, possession and spatial relations). These results hint at a trade-off: the stronger the relationship with geographic proximity, the weaker the relationship with cultural similarity (r = -0.53, p = 0.014). There was no statistically significant trade-off between cultural similarity and historical proximity (r = 0.27, p = 0.24).

To summarise, semantic alignment was to some extent predictable from cultural similarity and historical relationships between languages. Strikingly, semantic alignment between languages is better predicted by cultural similarity than by the geographic proximity of the populations who speak them.

Discussion

We computed semantic alignment for 1,010 meanings in 41 languages using distributed semantic vectors derived from multilingual natural language corpora. Comparing the structure of the resulting semantic spaces allowed us to measure at scale whether translation equivalents really mean the same thing in each language. To the extent that the vocabularies of different languages organize the world in similar ways — carving nature at its joints — their vocabularies are expected to converge on common categories and therefore ought to be highly alignable. In contrast, if different languages impose their own structure, carving joints into nature, word meanings may align to a more limited extent; a word and its translation equivalent may not mean quite the same thing and meanings that are easy to express in one language may not be easy to express in another.

We found semantic alignment to vary strongly with semantic domain (Figure 3). Words for common artifacts, actions, and natural kinds – meanings that ought to be highly aligned on a strong universalist account – were found to have only intermediate alignments. Contributing to this lower alignment were cross-linguistic differences in word frequency, differences in semantic neighborhood densities, and differences in patterns of polysemy.[15,34,82]

We observed highest semantic alignment in domains characterized by high internal structure: numbers, temporal terms, and kinship terms. This result suggests that the structure of e.g., a base-10 number system, a 12-month calendar — products of cultural evolution whose acquisition may require experience with language[83,84] — constrains semantic relationships among words like "five" and "ten", "month" and "year" in a very similar way in different languages. These high alignments may reflect a universal basis for representing these concepts,[85] but the fact that alignment for kinship and temporal terms was further predicted by nonlinguistic measures of cultural similarity speaks to the influence of culture on linguistic structure in these highly aligned domains. Though alignment of number words was not predicted by overall cultural similarity, alignment patterns of specific numerals for different languages were strongly influenced by the linguistic formation of these numerals (e.g., whether the word for 12 is atomic (as in English) or is a composite form (e.g., ten+two, as in Bulgarian), see SI 4.2.

It may be tempting to ascribe our finding that words denoting common natural kinds and other everyday meanings have only intermediate alignment to random noise or other inadequacies of our method. However, the extent of such (mis)alignment between different languages was not random, but predictable from estimates of historic, geographic, and cultural relatedness. Languages with greater shared history, languages geographically closer, and languages spoken by more similar cultures (as estimated from independent sources) had greater semantic alignment. This result shows that automatically derived natural language semantics contain a strong signal of cultural and historical processes.

Our reliance on corpus-derived semantic representations has limitations. Human semantic representations encode many relationships not present in semantics learned from word use alone.[86,87] Semantic knowledge automatically derived from corpora reflects only information contained in language and therefore under-represents what people learn about the meaning of words from direct interactions with the world. For example, the meaning of “dog” in distributional semantic terms is derived from the contexts in which the word occurs, and includes people’s direct experiences with dogs only to the extent that they are reflected in language. Although this makes distributional semantics an incomplete account of word meanings, distributional semantic analyses like ours can be viewed as a conservative estimate of cross-linguistic differences in word meaning: differences not reflected in language use are likely to only lower estimates of semantic alignment.

Our finding that many words denoting natural kinds and common, concrete meanings only show intermediate alignment between languages, combined with the finding that this alignment is related to cultural, historical, and geographic factors, conflicts with conclusions from some recent work[34,82] where semantic alignment was operationalized in terms of similarity of polysemy networks (i.e., translation equivalents are similar to the extent that they colexify in the same way). For example, Youn et al.[34] reported that polysemy-based alignment between geographically proximate or culturally similar languages was no greater than between randomly selected languages — "consistent with the hypothesis that cultural and environmental factors have little statistically significant effect on the semantic network of the [concepts] studied here". This absence of an effect was interpreted as supporting the universalist view. Our view is that analyses of polysemy networks are extremely valuable in that they help us understand how senses of words change over time. This process of change may indeed be very similar across languages.[88]

However, historical regularities of sense formation do not necessarily mean that translation equivalents with similar polysemy networks mean the same thing in each language, e.g., some of the senses may be much more frequent in one language or another resulting in less alignment than polysemy networks suggest. Conversely, translation equivalents with different polysemy networks do not necessarily mean very different things, e.g., some of the attested senses may not be in current use. Accordingly, our analyses reveal that polysemy-based alignment, while positively correlated with our alignment measure based on distributional semantics, was only weakly related to human translatability judgments (e.g., accounting for 6.5 times less variance in English-Dutch translatability human norms[24]), and 2.8 times less variance in accounting for cross-linguistic differences in consistency of picture naming[57]), see SI 5.3). These results suggest that of the two approaches to alignment, measures based on distributional semantics may more closely reflect differences in how words are actually used.

We were able to reproduce the absence of a relationship between polysemy-based alignment and geographical and cultural factors (SI 4.1), suggesting that the difference between the current results and previous findings does not stem exclusively from differences between the sample of concepts and languages analyzed, or from differences in how cultural similarity is operationalised. Our measure of alignment based on distributional models of human semantic representations was strongly associated with geographic and cultural proximity — a relationship supporting the relative over the universalist position.

We view our work as an early attempt to quantify semantic alignment at scale using distributional semantics. Advances in machine learning, such as new methods for unsupervised alignment of vector spaces,[89] and contextual word embeddings[90] - are likely to help scale this approach even further, and address some of its limitations. Although we were able to use several datasets to validate our alignments against human data,[24,26,57] and to verify that the results are robust to changes in training corpora (see SI 1.2.2), we recognize the need for additional validation using translatability ratings from multilingual subjects. Our alignment values can be used to compile stimulus lists for these studies in ways that maximize informativeness, e.g., by strategically choosing words that are predicted to have high and low translatability.

Our results do not fully fit into either the universalist or relative perspectives. The ranking of semantic domains by their alignment (Fig. 3) has unexpected elements when viewed through either the relative or the universalist lens. The finding that numerals, time, kinship, and sense words have relatively high alignment may be viewed as supporting the idea that these word meanings derive from universal cognitive and perceptual biases. However, the finding that alignment is uncorrelated with concreteness and that some of the most concrete domains have relatively low alignment is unexpected on the universalist view, as are our findings that alignment of even relatively aligned domains such as kinship and temporal terms is strongly influenced by cultural similarity.

Our findings do not rule out the existence of universal semantic primitives into which many common words can be decomposed,[91,92] but see.[93,94] We think progress in this direction is likely to come from large-scale efforts focusing on aligning languages using multilingual embeddings of words and larger verbal constructions derived from naturalistic language corpora.

Methods

Data

The primary dataset we examined (further details in SI 1.4.2) is a subset of the intersection of NorthEuraLex (NEL) and fast-text word embeddings trained on Wikipedia, filtered to exclude: any languages whose Wikipedia data do not exceed a small set of quality criteria and any concepts not present in at least 20 languages. These filtering criteria, described in more detail in SI 1.2.3, did not have a significant impact on our conclusions. SI 2 provides details for all of the analyses reported in main text, including the statistical tests we employed, as well as the number of languages, language-pairs, concepts, and language families (these details varied from analysis to analysis because not all of the language pairs whose alignment we calculated were available in the all of the data sources we examined).

Of note, the 39 languages used in relating alignment to cultural similarity, geography, and language history (SI 4) was not a strict subset of the 41 languages used in the main analyses because restricting the sample to the languages that passed our filtering criteria reduced the overlap to 20 languages. To account for potentially low concept coverage in some of these languages, we included number of concepts as a covariate in the models. Repeating the analyses on the 20-language subset yields the same conclusions.

We mapped concepts listed in NEL to entries in the Intercontinental Dictionary Series (IDS) using the Concepticon.[95] The IDS is structured into chapters which we used to assign each of the NEL concepts to a semantic domain. The full list of semantic domains and their mapping to NEL concepts is provided in SI 1.4.2.

Algorithm

SI 1.1 provides a formal description of the procedure we used to calculate alignment for word-pairs, concepts, language-pairs, and semantic domains. Aggregate cross-linguistic alignment (at the concept- and domain-level) reflects simple averages taken over word-pair level alignment in all language pairs for which relevant data was available. All analyses reported in the main text used neighbor search depth $k = 100$.

Replications

We replicated our main findings using alternative word embeddings and translation sets (including a much larger set of 20,000 translation equivalents available for a smaller number of languages[96]). In SI 1.2., we show strong correlations between our primary analyses and these replications, at the level of word-pairs, concepts, language-pairs, and semantic domains. These replications justify the decision to treat alignment among the NEL concepts calculated from Wikipedia-trained embeddings as our primary dataset.

We also analyzed: deliberately corrupted corpora to establish baseline rates of alignment (SI 1.3.1); alignment between two embeddings models of a single language (English) trained on different corpora (SI 1.3.2); alignment in embeddings spaces trained on lemmatized corpora (SI 1.2.5); alignment using different numbers of semantic neighbours (the only free parameter of our algorithm, SI 1.2.2); and alignment calculated using alternative measures of structural similarity in vector spaces (SI 1.1.4).

Cultural Similarity

The measure of cultural similarity was based on cultural traits from the Ethnographic Atlas as linked to languages in D-PLACE.[67] Missing values were imputed by multiple imputation using classification and regression trees,[97] including language family as a conditioning factor. During testing, this method imputed the correct value for held-out data 74% of the time, compared to a baseline of imputation by random choice of 19%. Cultural distances between two language groups were calculated as the average Gower distances between traits in 100 imputed sets. See SI 4.1 for more details.

Historic and Geographic Relationships

Historic proximity was measured using patristic distances in a phylogenetic tree of 19 Indo-European languages.[80] Geographic proximity was measured as the great circle distance between the cultural centres of each language as defined in Glottolog.[81] See SI 4.1 for further details.

Data Availability

Data and reproducible analyses are available at `https://osf.io/tngba/`

Code Availability

Code to implement the alignment algorithm is available at `https://osf.io/tngba/`

References

[1] Gleitman, L. & Fisher, C. Universal aspects of word learning. In McGilvray, J. (ed.) *The Cambridge Companion to Chomsky*, 123–142 (Cambridge University Press, New York, NY, 2005).

[2] Snedeker, J. & Gleitman, L. Why is it hard to label our concepts? In Hall, D. G. & Waxman, S. (eds.) *Weaving a Lexicon*, 257–294 (The MIT Press, Cambridge, MA., 2004), illustrated edition edn.

[3] Pinker, S. *The Language Instinct* (Harper Collins, New York, 1994).

[4] Berlin, B. & Kay, P. *Basic Color Terms: Their Universality and Evolution* (University of California Press, 1969).

[5] Li, P. & Gleitman, L. Turning the tables: Language and spatial reasoning. *Cognition* **83**, 265–294 (2002).

[6] Evans, N. & Levinson, S. C. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences* **32**, 429–448 (2009).

[7] Lupyan, G. The centrality of language in human cognition. *Language Learning* **66**, 516–553 (2016).

[8] Lupyan, G. & Dale, R. Why are there different languages? The role of adaptation in linguistic diversity. *Trends in Cognitive Sciences* **20**, 649–660 (2016).

[9] Davidson, D. On the Very Idea of a Conceptual Scheme. *Proceedings and Addresses of the American Philosophical Association* **47**, 5–20 (1973).

[10] Lupyan, G. & Zettersten, M. Does vocabulary help structure the mind? In *Minnesota Symposia on Child Psychology. Vol 40*.

[11] Whorf, B. *Language, Thought, and Reality* (MIT Press, Cambridge, MA, 1956).

[12] Zgusta, L. *Manual of lexicography* (Mouton, 1971).

[13] Haspelmath, M. Lexical borrowing: Concepts and issues. *Loanwords in the world's languages: A comparative handbook* 35–54 (2009).

[14] Myers-Scotton, C. *Contact linguistics: Bilingual encounters and grammatical outcomes* (Oxford University Press on Demand, 2002).

[15] Xu, Y., Duong, K., Malt, B. C., Jiang, S. & Srinivasan, M. Conceptual relations predict colexification across languages. *Cognition* **201**, 104280 (2020).

[16] Regier, T., Carstensen, A. & Kemp, C. Languages support efficient communication about the environment: Words for snow revisited. *PloS one* **11**, e0151138 (2016).

[17] Winter, B., Perlman, M. & Majid, A. Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition* **179**, 213–220 (2018).

[18] Majid, A. *et al.* Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences* **115**, 11369–11376 (2018).

[19] San Roque, L., Kendrick, K. H., Norcliffe, E. & Majid, A. Universal meaning extensions of perception verbs are grounded in interaction. *Cognitive Linguistics* **29**, 371–406 (2018).

[20] Svensén, B. *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making* (Cambridge University Press), 1 edition edn.

[21] Cuyckens, H., Dirven, R. & Taylor, J. R. *Cognitive approaches to lexical semantics*, vol. 23 (Walter de Gruyter, 2009).

[22] Barnett, G. A. Bilingual semantic organization: A multidimensional analysis. *Journal of Cross-Cultural Psychology* **8**, 315–330 (1977).

[23] Moldovan, C. D., Sánchez-Casas, R., Demestre, J. & Ferré, P. Interference effects as a function of semantic similarity in the translation recognition task in bilinguals of catalan and spanish. *Psicologica: International Journal of Methodology and Experimental Psychology* **33**, 77–110 (2012).

[24] Tokowicz, N., Kroll, J. F., De Groot, A. M. & Van Hell, J. G. Number-of-translation norms for dutch—english translation pairs: A new tool for examining language production. *Behavior Research Methods, Instruments, & Computers* **34**, 435–451 (2002).

[25] Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M. & Baayen, H. How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and language* **62**, 284–301 (2010).

[26] Allen, D. & Conklin, K. Cross-linguistic similarity norms for Japanese–English translation equivalents. *Behavior Research Methods* **46**, 540–563 (2014).

[27] Bradley, M. M. & Lang, P. J. Affective norms for english words (ANEW): Instruction manual and affective ratings (1999). Technical report: Center for Research in Psychophysiology, University of Florida.

[28] Fairfield, B., Ambrosini, E., Mammarella, N. & Montefinese, M. Affective norms for italian words in older adults: Age differences in ratings of valence, arousal and dominance **12**, e0169472. Publisher: Public Library of Science.

[29] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. Introduction to wordnet: An on-line lexical database. *International journal of lexicography* **3**, 235–244 (1990).

[30] Sigman, M. & Cecchi, G. A. Global organization of the wordnet lexicon. *Proceedings of the National Academy of Sciences* **99**, 1742–1747 (2002).

[31] Majid, A., Jordan, F. & Dunn, M. Semantic systems in closely related languages. *Language Sciences* 1–18 (2015).

[32] Calude, A. S. & Verkerk, A. The typology and diachrony of higher numerals in indo-european: a phylogenetic comparative study. *Journal of Language Evolution* **1**, 91–108 (2016).

[33] Verkerk, A. Where do all the motion verbs come from?: The speed of development of manner verbs and path verbs in indo-european. *Diachronica* **32**, 69–104 (2015).

[34] Youn, H. *et al.* On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences* **113**, 1766–1771 (2016).

[35] Vivas, L., Montefinese, M., Bolognesi, M. & Vivas, J. Core features: measures and characterization for different languages. *Cognitive Processing* 1–17 (2020).

[36] Jackson, J. C. *et al.* Emotion semantics show both cultural variation and universal structure. *Science* **366**, 1517–1522.

[37] Firth, J. R. *Papers in Linguistics 1934-1951* (London: Oxford University Press, 1957).

[38] Lund, K. & Burgess, C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers* **28**, 203–208 (1996).

[39] Elman, J. An alternative view of the mental lexicon. *Trends in Cognitive Sciences* **8**, 301–306 (2004).

[40] Landauer, T. K. & Dumais, S. T. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review* **104**, 211–40 (1997).

[41] Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[42] Baroni, M., Dinu, G. & Kruszewski, G. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 1, 238–247 (2014).

[43] Baroni, M. & Lenci, A. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* **36**, 673–721 (2010).

[44] Hollis, G. & Westbury, C. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic bulletin & review* **23**, 1744–1756 (2016).

[45] Nematzadeh, A., Meylan, S. C. & Griffiths, T. L. Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *Proceedings of the 39th annual meeting of the cognitive science society*, 859–864 (2017).

[46] Levy, O. & Goldberg, Y. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, 2177–2185 (2014).

[47] Hill, F., Reichart, R. & Korhonen, A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456* (2014). `1408.3456`.

[48] Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).

[49] Boleda, G. Distributional semantics and linguistic theory. *Annual Review of Linguistics* **6**, 213–234 (2020).

[50] Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95 (2019).

[51] De Deyne, S., Perfors, A. & Navarro, D. J. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1861–1870 (2016).

[52] Šipka, D. *Lexical Conflict: Theory and Practice* (Cambridge University Press, 2015).

[53] Dellert, J. *et al.* NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation* **54**, 273–301 (2020).

[54] Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* **5**, 135–146 (2017).

[55] Lison, P. & Tiedemann, J. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2016)* (2016).

[56] Grave, E., Bojanowski, P., Gupta, P., Joulin, A. & Mikolov, T. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)* (2018).

[57] Duñabeitia, J. A. *et al.* MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology (2006)* **71**, 808–816 (2018).

[58] Brysbaert, M., Warriner, A. B. & Kuperman, V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* **46**, 904–911 (2014).

[59] Jones, D. Human kinship, from conceptual structure to grammar. *Behavioral and Brain Sciences* **33**, 367 (2010).

[60] Kemp, C. & Regier, T. Kinship Categories Across Languages Reflect General Communicative Principles. *Science* **336**, 1049–1054 (2012).

[61] Givón, T. On the development of the numeral 'one' as an indefinite marker. *Folia Linguistica Historica* **15**, 35–54 (1981).

[62] Rzymski, C. *et al.* The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data* **7**, 1–12 (2020).

[63] Dehaene, S. & Mehler, J. Cross-linguistic regularities in the frequency of number words. *Cognition* **43**, 1–29 (1992).

[64] Vecchi, E. M., Baroni, M. & Zamparelli, R. Linear maps of the impossible: capturing semantic anomalies in distributional space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, 1–9 (2011).

[65] Speer, R., Chin, J., Lin, A., Jewett, S. & Nathan, L. Luminosoinsight/wordfreq: v2.2 (2018). URL `https://doi.org/10.5281/zenodo.1443582`.

[66] Pagel, M., Atkinson, Q. D. & Meade, A. Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature* **449**, 717–720 (2007).

[67] Kirby, K. R. *et al.* D-place: A global database of cultural, linguistic and environmental diversity. *PloS one* **11**, e0158391 (2016).

[68] Murdock, G. P. & Provost, C. Factors in the division of labor by sex: A cross-cultural analysis. *Ethnology* **12**, 203–225 (1973).

[69] Sellen, D. W. & Smay, D. B. Relationships between subsistence and age at weaning in "preindustrial" societies. *Human Nature* **12**, 47–87 (2001).

[70] Li, P. & Gleitman, L. Turning the tables: Language and spatial reasoning. *Cognition* **83**, 265–294 (2002).

[71] Apostolou, M. Bridewealth as an instrument of male parental control over mating: Evidence from the standard cross-cultural sample. *Journal of Evolutionary Psychology* **8**, 205–216 (2010).

[72] Meggers, B. J. Environmental limitation on the development of culture 1. *American anthropologist* **56**, 801–824 (1954).

[73] Peoples, H. C. & Marlowe, F. W. Subsistence and the evolution of religion. *Human Nature* **23**, 253–269 (2012).

[74] Botero, C. A. *et al.* The ecology of religious beliefs. *Proceedings of the National Academy of Sciences* **111**, 16784–16789 (2014).

[75] Gavin, M. C. *et al.* The global geography of human subsistence. *Royal Society Open Science* **5**, 171897 (2018).

[76] Martin, M. K. & Voorhies, B. *Female of the Species* (Columbia University Press, New York, 1975). Accessed on: 2019-01-18.

[77] Goodenough, W. H. Basic economy and community. *Behavior Science Notes* **4**, 291–298 (1969).

[78] Lakoff, G., Espenson, J. & Schwartz, A. The mastermetaphor list, 2nd ed. TechnicalReport, University of California Berkeley.

[79] Wiseman, R. Interpreting ancient social organization: Conceptual metaphors and image schemas. *Time and Mind* **8**, 159–190 (2015).

[80] Bouckaert, R. *et al.* Mapping the origins and expansion of the indo-european language family. *Science* **337**, 957–960 (2012).

[81] Hammarström, H., Forkel, R. & Haspelmath, M. clld/glottolog: Glottolog database 3.3. Jena: Max Planck Institute for the Science of Human History (2018).

[82] Srinivasan, M. & Rabagliati, H. How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua* **157**, 124–152 (2015).

[83] Gordon, P. Numerical cognition without words: Evidence from amazonia **306**, 496–499.

[84] Tillman, K. A. & Barner, D. Learning the language of time: Children's acquisition of duration words. *Cognitive psychology* **78**, 57–77 (2015).

[85] Gelman, R. & Butterworth, B. Number and language: how are they related? **9**, 6–10.

[86] Chen, D., Peterson, J. C. & Griffiths, T. L. Evaluating vector-space models of analogy. *arXiv:1705.04416 [cs]* (2017). `1705.04416`.

[87] Huebner, P. A. & Willits, J. A. Structured Semantic Knowledge Can Emerge Automatically from Predicting Word Sequences in Child-Directed Speech. *Frontiers in Psychology* **9** (2018).

[88] Ramiro, C., Srinivasan, M., Malt, B. C. & Xu, Y. Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences* **115**, 2323–2328 (2018).

[89] Grave, E., Joulin, A. & Berthet, Q. Unsupervised alignment of embeddings with wasserstein procrustes. *arXiv preprint arXiv:1805.11222* (2018).

[90] Peters, M. E. *et al.* Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[91] Wierzbicka, A. *Semantics: Primes and universals* (Oxford University Press, UK).

[92] Goddard, C. & Wierzbicka, A. *Meaning and Universal Grammar: Theory and Empirical Findings* (John Benjamins Publishing).

[93] Aitchison, J. *Words in the Mind: An Introduction to the Mental Lexicon* (Wiley-Blackwell), 4th edn.

[94] Matthewson, L. Is the meta-language really natural? *Theoretical Linguistics* **29**, 263–274.

[95] List, J. M., Greenhill, S., Rzymski, C., Schweikhard, N. & Forkel, R. (eds.) *Concepticon 2.0* (Max Planck Institute for the Science of Human History, Jena, 2019).

[96] Conneau, A., Lample, G., Ranzato, M., Denoyer, L. & Jégou, H. Word translation without parallel data. *arXiv preprint arXiv:1710.04087* (2017).

[97] Buuren, S. v. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. *Journal of statistical software* 1–68 (2010).

[98] Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution* **23**, 254–267 (2005).

Acknowledgments

BT & GL acknowledge support from LEVINSON fellowships at the Max Planck Institute for Psycholinguistics. SGR was partially supported by a Leverhulme early career fellowship (ECF-2016-435). This was partially funded by NSF-PAC 1734260 to GL. Funders had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript. The authors wish to thank Johannes Dellert and Asifa Majid.

Author Contributions

BT, SGR, GL designed the research, collected and analyzed data, and contributed to the writing of the manuscript.

Competing Interests

The authors declare no competing interests.

Figure Legends

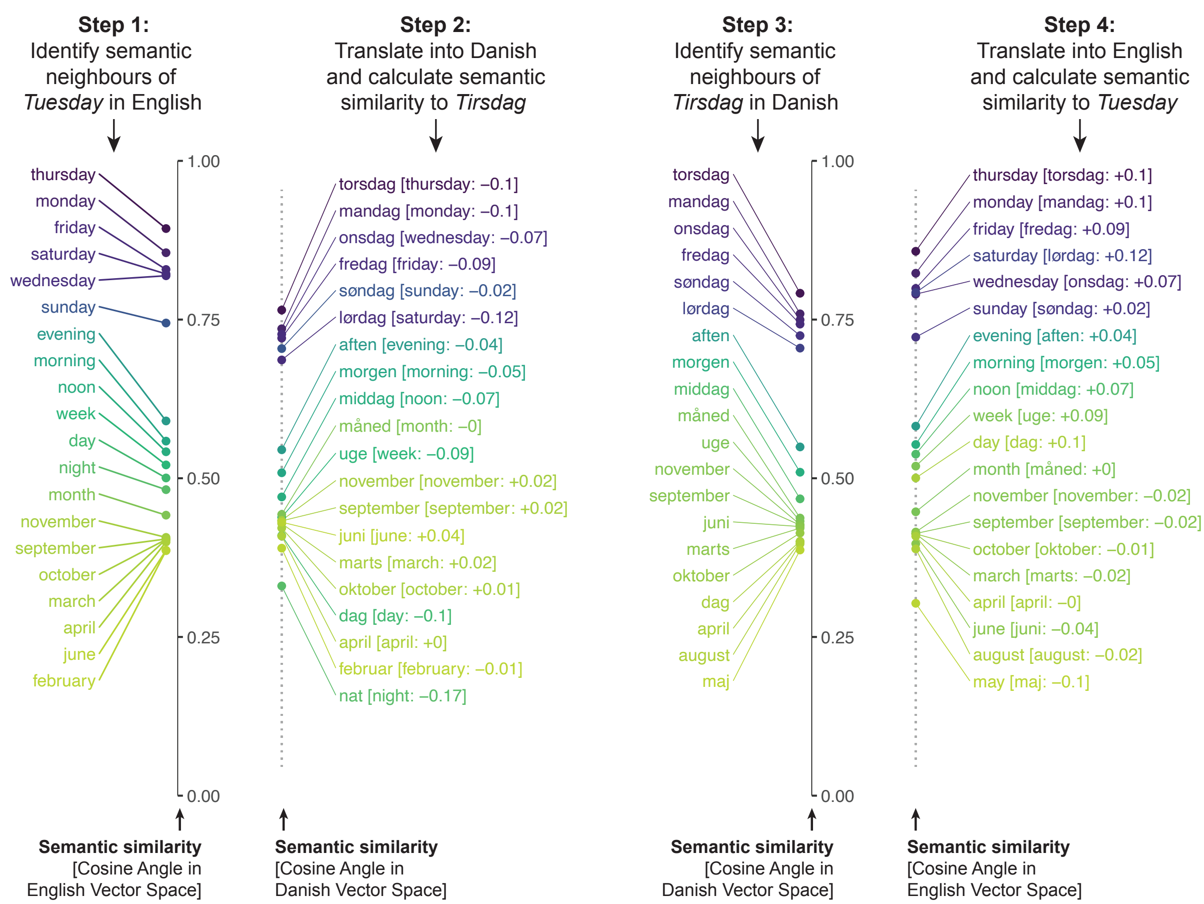

Figure 1: **High alignment between English ("Tuesday") and Danish ("Tirsdag")**. A schematic of the algorithm for computing semantic alignment. Colour denotes semantic similarity in the first language; similar colour ordering on both sides of the plot indicates high alignment.

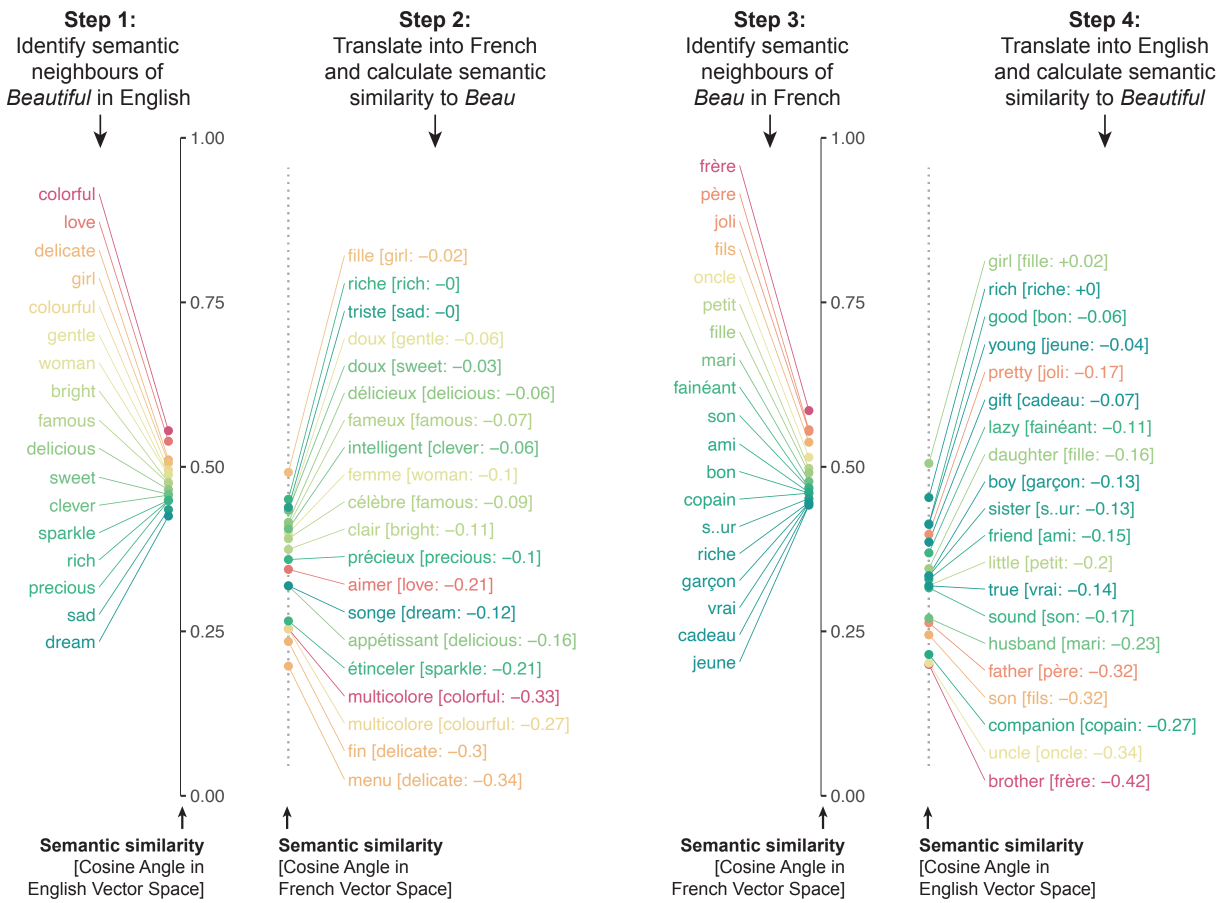

Figure 2: **Low alignment between English ("beautiful") and French ("beau")** A schematic of the algorithm for computing semantic alignment. Colour denotes semantic similarity in the first language; perturbed colour ordering indicates low alignment.

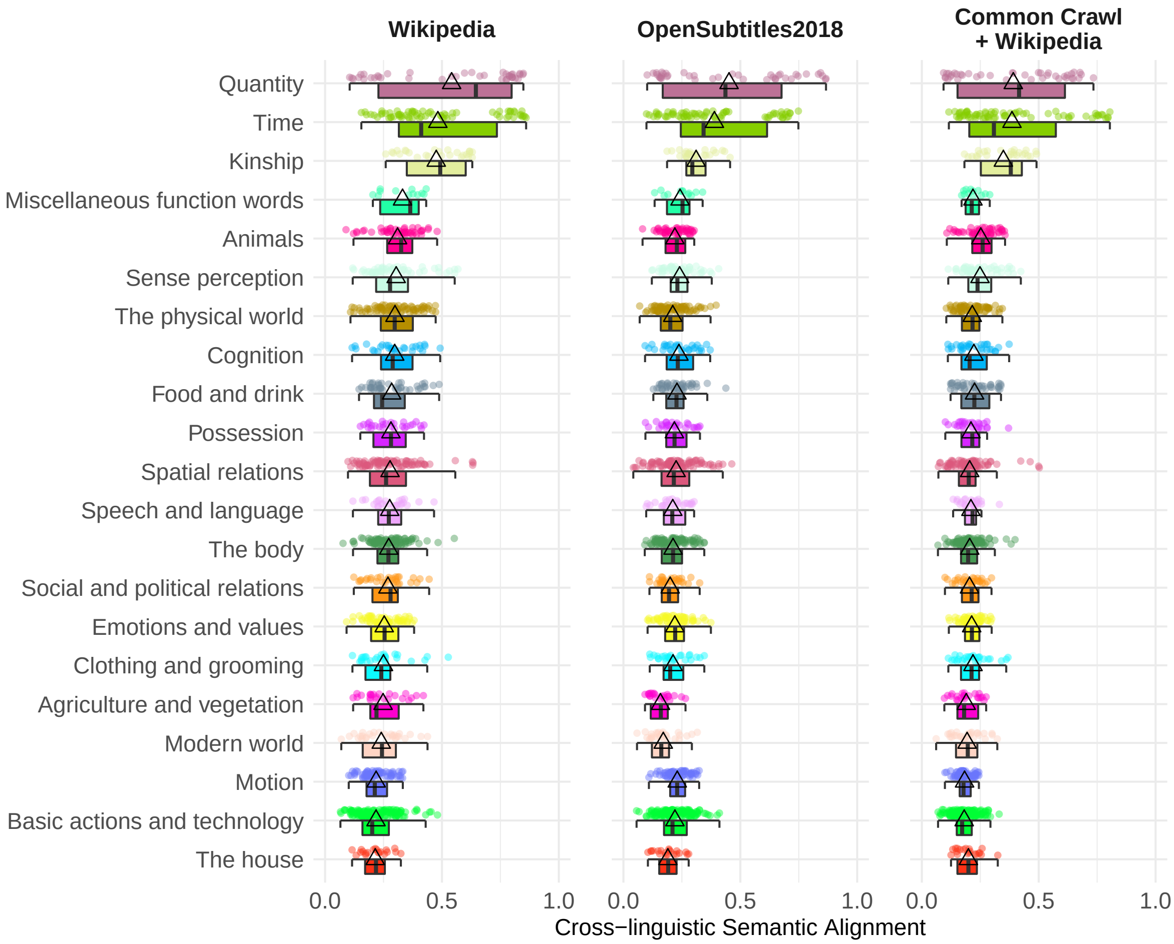

Figure 3: **Semantic Alignment of 21 Semantic Domains** Semantic domains ranked according to mean cross-linguistic semantic alignment, as computed from word embeddings induced from Wikipedia (left), OpenSubtitles 2018 (middle), and Wikipedia & Common Crawl (right). Each point shows semantic alignment for a unique pair of languages averaged over all translation pairs in the relevant semantic domain. Boxplots here and throughout show: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x inter-quartile range. Triangles show mean alignment.

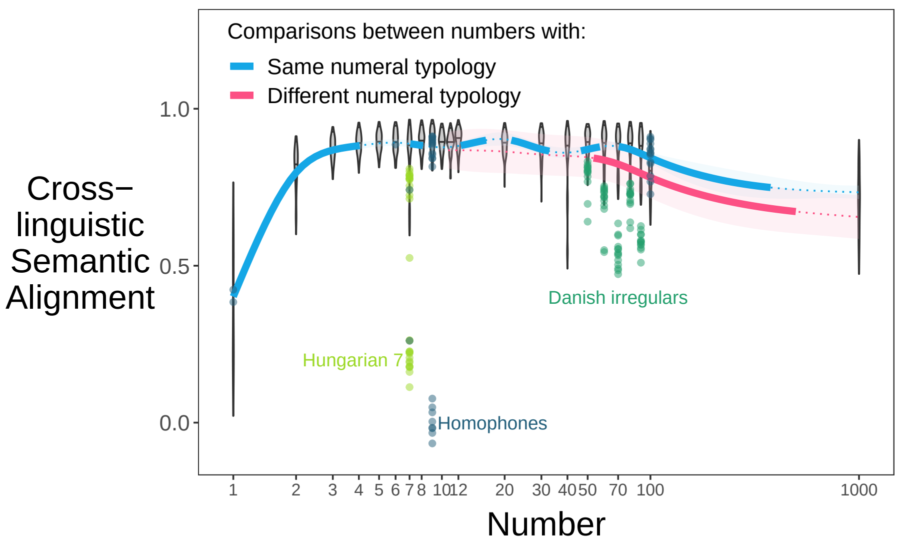

Figure 4: **Semantic alignment of number words** Semantic alignment between 16 languages for 22 number words. Each violin plot shows the distribution of alignment values. GAM lines are plotted for comparisons between numbers with the same numeral typology (blue) and different numeral typology (pink). The ribbons show 95% confidence interval around the mean and solid lines indicate areas of significant increase or decrease. Therefore, the main difference in numeral typology applies to numbers above 40. Various outliers belong to three groups: comparisons with Hungarian 7, words with alternative meanings (e.g. French "neuf" meaning '9' or 'new'), and Danish irregular numbers.

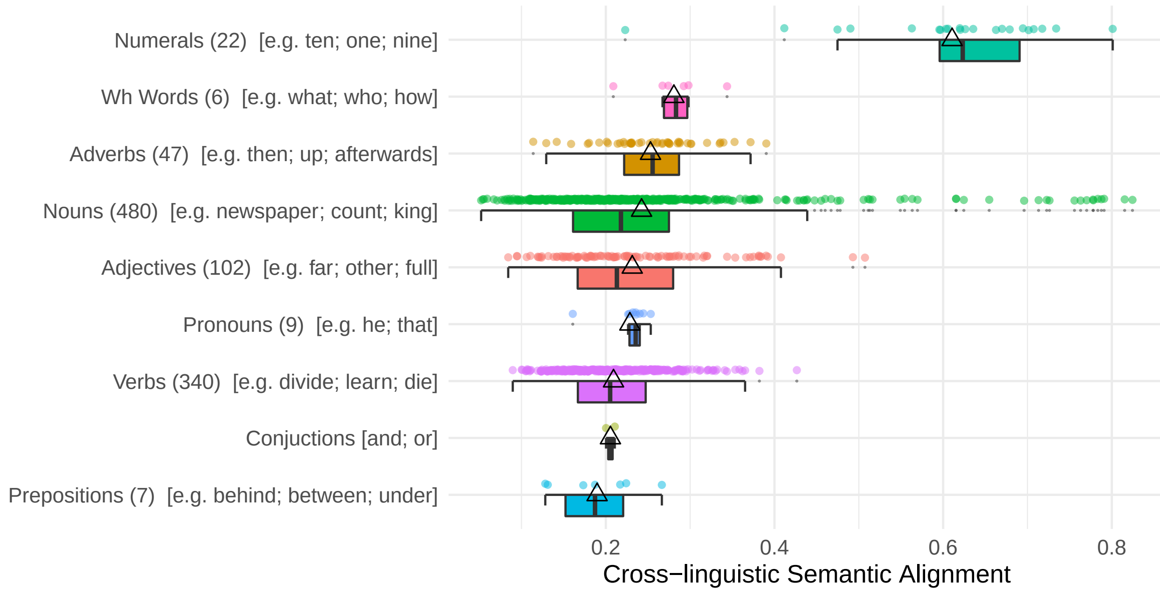

Figure 5: **Semantic Alignment by Part of Speech** Numerals were most strongly aligned across languages, followed by Wh-words and adverbs. Prepositions were least aligned. Each point is the average alignment of one concept across all pairs of languages. Triangles show mean cross-linguistic alignment.

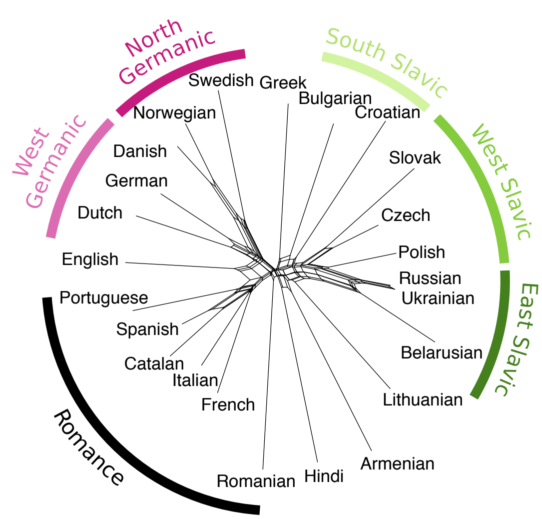

Figure 6: **Semantic distances for Indo-European languages** Semantic distances between languages visualised as a neighbour-net (using Splitstree,,[98] see SI 4.3). Distances are represented along the shortest path between language nodes. The semantic distances reflect established historical relationships, as shown by the labelling of the major sub-branches according to Glottolog.[81] Conflicting signal shows up as parallel lines. For example, English shows conflicting signal between Germanic and Romance, which reflects its mixed history